The Report Committee for Ryan Christian Golden
Certifies that this is the approved version of the following report:

NewsFerret:
Supporting Identity Risk Identification and Analysis Through Text Mining of News Stories

APPROVED BY
SUPERVISING COMMITTEE:

Supervisor:

Kathleen Suzanne Barber

Christine Julien

NewsFerret:
Supporting Identity Risk Identification and Analysis Through Text Mining of News Stories

by

Ryan Christian Golden, B.A.

Report

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Master of Science in Engineering

The University of Texas at Austin

May 2013

Dedication

I dedicate this to my wife and two sons, who lovingly support my many identities.

Acknowledgements

I offer my sincere gratitude to my supervisor, Dr. Suzanne Barber, who helped guide me to this topic, and whose founding and development of the Center for Identity at the University of Texas at Austin has made it possible for my own and others' research to continue in the important and still-emerging field of identity. Thanks to Dr. Christine Julien, an inspiring lecturer, whose course taught me—an English major in a former life—the all-important skills of how to read and author engineering papers. Suratna Budalakoti and other lab members at the Center for Identity suggested valuable tools and feedback during my research. I thank all the professors, teaching assistants, staff, and administration at the Option III Software Engineering Program for putting together a life-changing degree program for working professionals. Their supportive and innovative program made it possible for me to continue and revitalize my education and research in the superb environment that is the University of Texas at Austin. I am grateful to Reza B'Far for his assistance, financial and otherwise, that enabled me to pursue my degree and professional career simultaneously.

I thank my mother and father for their early role in the process of my education and for their constant love and support. Finally, I thank my wife for supporting me with the love, care, and hard work I needed to get through the challenging process of completing a graduate degree while also working full-time and raising two amazing, rambunctious young boys.

Abstract

NewsFerret:
Supporting Identity Risk Identification and Analysis Through Text Mining of News Stories

Ryan Christian Golden, MSE

The University of Texas at Austin, 2013

Supervisor: Kathleen Suzanne Barber

Individuals, organizations, and devices are now interconnected to an unprecedented degree. This has forced identity risk analysts to redefine what "identity" means in such a context, and to explore new techniques for analyzing an ever expanding threat context. Major hurdles to modeling in this field include the inherent lack of publicly available data due to privacy and safety concerns, as well as the unstructured nature of incident reports. To address this, this report develops a system for strengthening an identity risk model using the text mining of news stories. The system—called *NewsFerret*—collects and analyzes news stories on the topic of identity theft, establishes semantic relatedness measures between identity concept pairs, and supports analysis of those measures through reports, visualizations, and relevant news stories. Evaluating the resulting analytical models shows where the system is effective in assisting the risk analyst to expand and validate identity risk models.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

Incidents of identity theft, fraud, and abuse continue to affect large numbers of individuals, small businesses, corporations, and government agencies in the U.S. and around the world. Recent surveys have estimated that between 2004 and 2011, more than 10 million U.S. adults have been affected each year, with 11.6 million victimized in 2011 at a monetary cost of $18 billion [1]. To help understand, assess, and control these identity threats, the Center for Identity (CID) at the University of Texas is developing systems to assist with identity risk modeling. These systems will provide a framework in which one can study the structure of an identity, its areas of vulnerability, and the estimated costs to personal, financial, or national security should elements of its structure be compromised. The Center is also seeking to create an analytical repository of known identity threats and counter measures, structured in a form suitable for analysis [2]. A crucial problem these systems face is how to acquire and process enough data on identity threats in order to validate models of identity risk. This report illustrates and evaluates a practical solution to this problem—a system called *NewsFerret*—that collects, models, and analyzes large numbers of news stories on identity threats.

IDENTITY RISK DOMAIN

To understand the value of *NewsFerret*, one must first understand the identity risk domain. This section introduces the scenarios, concepts, and objectives of the identity risk domain at a high level. These are further illustrated by a concrete example in the next section.

Organizations employ the technique of identity risk modeling to combat and manage identity theft, fraud, and abuse threats. Some scenarios where identity risk modeling would be important include:

- A company or government agency must choose how to focus resources in response to an identity theft or data breach incident.
- Law enforcement conducting forensics on an identity theft must identify potential suspects or attack vectors.
- A software or network architect must design a system that will store, authenticate, and use identifying information.

What organizations wish to protect in each of the above scenarios are the *identity attributes*. Identity attributes are the elements that are intrinsic to an identity or set of identities within a particular domain and that have value. As an example, consider all the persons interacting with an e-commerce web site. The identity attributes for these persons might include:

- User Login
- E-mail Address
- Credit Card Number
- Mailing Address
- Password

The *identity ecosystem* is the collection of all such identity attributes plus all the relations between those attributes. Example relations might include:

- A password grants access to a bank account
- A social security number authorizes a tax return

The actors, actions, and resources that serve to steal, defraud or abuse elements of the identity ecosystem are what make up the *threat context*. For example, a hacker may use a bogus credit card number generator to gain access to an Amazon account. An identity thief may steal mail from mailboxes in order to forge a job application. The combination of all possible threats makes up the threat context.

The abstract region where an identity ecosystem and a threat context overlap is referred to as the *identity risk* of a domain. Organizations that deal with identities naturally wish to reduce identity risk so as to reduce cost and liability to their business. Where risk cannot be eliminated, organizations seek to understand and control it through a set of steps that make use of an *identity risk model* (an example of which is described in the next section). These steps are explained in further detail in the Requirements section:

- Identify risk
- Analyze risk
- Assess risk
- Manage risk

To be successful in these steps requires a valid risk model; in other words, the risk model should represent as faithfully as possible the underlying, real-world, *semantic structure* of the identities and threats being modeled. The effort to capture this real-world structure through a model can be visualized as follows:

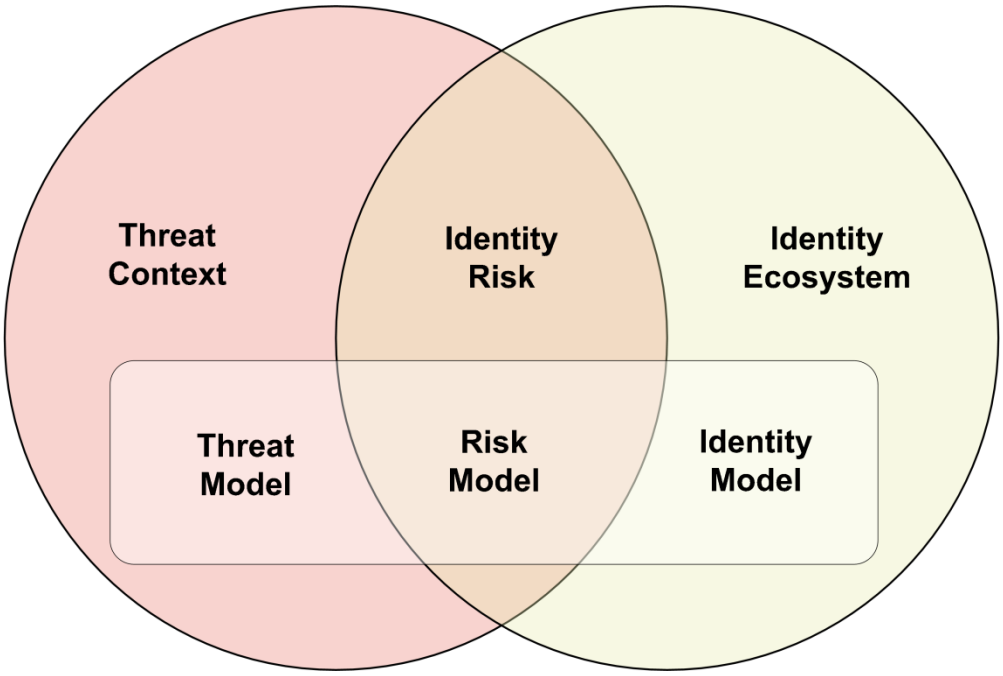

Figure 1: Modeling identity risk

As Figure 1 illustrates, a risk model typically captures only a portion of the real-world structure of the ecosystem and threat context. To increase the coverage and accuracy of an identity risk model, the analyst faces a challenge, since the true structures of the ecosystem and threat context are typically not fully known or understood without further research. This is where *NewsFerret* can assist.

EXAMPLE SCENARIO

To more concretely illustrate the process of identity risk modeling, the author introduces a fictitious example scenario and risk model. The risk model is revisited periodically in later sections to show how *NewsFerret* might help to evolve and strengthen the model.

The scenario is as follows. A number of credit card thefts have recently been reported by users of the Amazon web site. The attack vector is unknown, although

internal security teams have theorized that the hackers are exploiting a widely reported technique of calling customer support to first add a fake credit card number and then calling again to add a new e-mail address to the account [3]. This technique compromises the user account which leads the hacker to the user's credit card information. The technique requires the following identity attributes: user's name, billing address, and e-mail address. The resources the hacker requires include a fake credit card number generator and the customer support phone line.

Based on the theory above, the internal security team has created the following risk model:

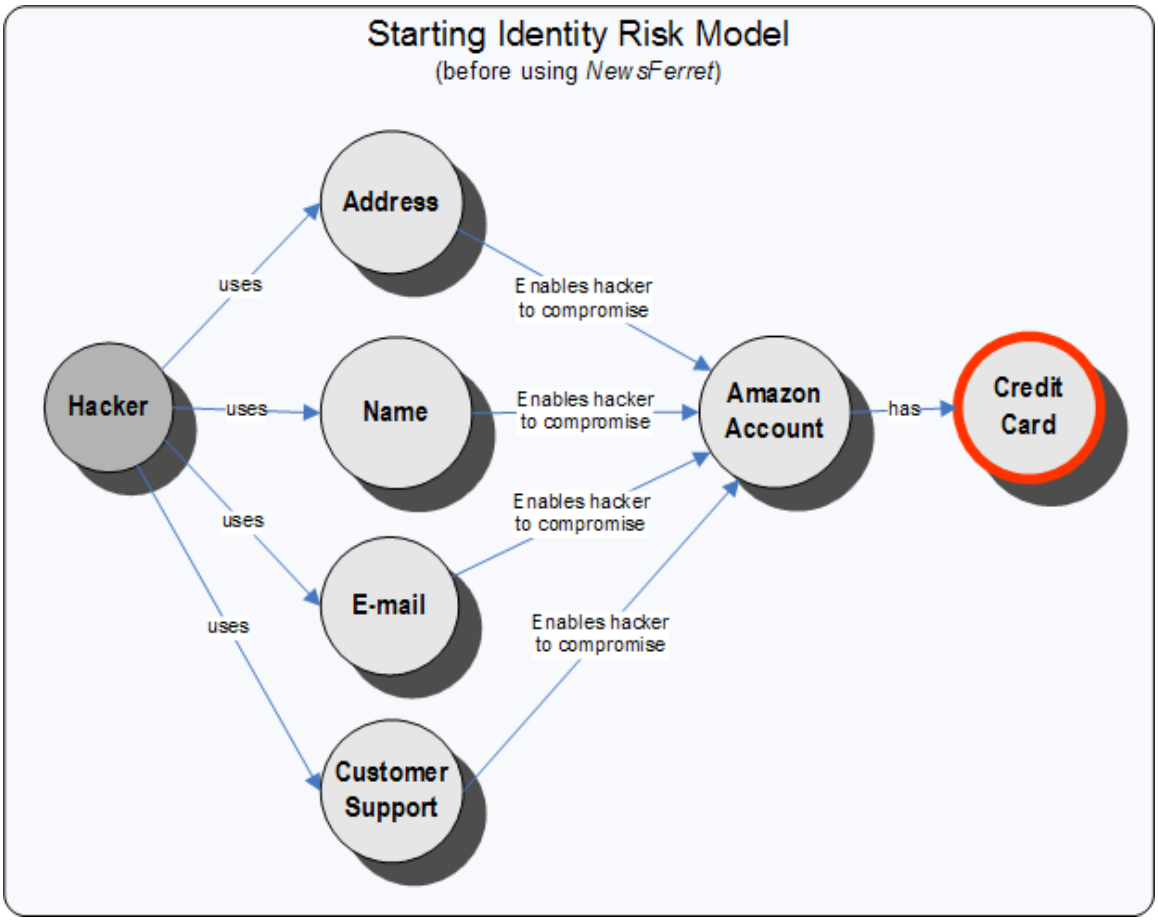

Figure 2: Example risk model before validation

Shaded elements indicate threat actors. The remaining nodes are identity attributes for an Amazon user. Red borders indicate a high value identity attribute. Labeled links between nodes indicate relations between identity attributes.

However, nobody is quite sure how accurate this model is. Nor can anyone put any value to the level of risk posed by individual elements in the model. This is because—aside from the number of reported credit card thefts—there is no data upon which to base any conclusions.

Consequently, as a security consultant specializing in identity risk modeling, you have been hired to lead a team of risk analysts to validate the client's identity risk model and answer the following questions:

- What are the likely attack vectors, i.e., the elements from which the attack originated?
- Are any elements missing in the above risk model?
- What elements pose the greatest risk, i.e., where should the client expend energy and resources to mitigate risk?

Your team now has all the pieces of information necessary to use *NewsFerret*: a description of the threat context (credit card information is being stolen), a description of the identity ecosystem (user name, credit card, e-mail, etc.), and a starting model. Subsequent sections of this paper revisit this example to illustrate how this information can be fed into *NewsFerret* to help validate this model and answer the above questions.

THE DATA PROBLEM

To validate risk models like the example model above, analytical tools must be backed by a sufficient amount of data to support an analyst's conclusions. One possible source for this data is a repository of structured incident records. However, such a repository requires data entry by a domain expert familiar with the structured incident format. The identity threat scenarios modeled in this way for the Center for Identity's incident repository—the Identity Threat Assessment and Prediction (ITAP) system)—are not yet sufficient in number. At the same time, there exists somewhat of an inverse problem. The global news media reports hundreds of identity incidents each month. Yet manual data entry of all of these reports into ITAP is not feasible due to their unstructured and voluminous nature. *NewsFerret* enables meaningful analysis given this dual problem:

1. Too little structured threat data (incident repositories)
2. Too much unstructured threat data (news stories)

To do this, *NewsFerret* realizes a design of a text mining system [4] that measures semantic relatedness between sets (e.g., pairs) of concepts present in the identity risk space. At a high level, the system works as follows: it collects news stories from the web based on a set of keyword-based topics (example keywords are "identity theft" and "identity fraud"); it analyzes the news stories using text mining techniques that include latent semantic analysis [5]; and it produces reports and analytical models suitable for identity risk identification, analysis, visualization and export to other analytical models.

The rest of this report is structured as follows. Chapter 2 places this report in the context of related work. Then, Chapter 3 details the requirements for identity risk model validation. Chapter 4 introduces the design of the *NewsFerret* system. Chapter 5 then analyzes the process and results of a month-long run of the system, demonstrates the analytical outputs of that system, and revisits the example scenario given above. Chapter 6 concludes with a summary of contributions, and Chapter 7 lays out ideas for future work.

Chapter 2: Related Work

NewsFerret resides within a context of related works in the areas of threat modeling, industry and government studies and initiatives, ongoing projects at the Center for Identity, and text mining studies.

THREAT MODELING

Although the threats of identity theft, fraud, and abuse are perhaps centuries old, the modern identity threat space has evolved rapidly in recent years as individuals, organizations, and devices have quickly become accustomed to high levels of connectivity to the rest of the world. Successive waves of innovative technologies—including the web, e-commerce, smart phones, tablets, social media, big data, and cloud services—have played a major part in this by rapidly establishing themselves as important and popular ways of interacting with and doing business in the world. Studies of smart phone adoption have shown that as of February 2012, 87% of American adults have cell phones, with 46% having smart phones [6]. As of March 2013, over 52% of Americans are Facebook users [7]. And in 2012, the number of internet-connected devices around the world grew to above 9 billion [8].

While many traditional threats to identity still persist (e.g., mail theft), increasing levels of connectedness and new technologies are exposing new inroads to malicious actors, driving a need to update existing definitions of identity and study a new landscape of identity risks. New identity attributes now exist and lay exposed, for example, *check-in location* [9]. Other well-established identity attributes like *credit card number* and *social security number* face new threats as they are digitized and replicated in big data systems around the world[8]. Entirely new populations of identities have been created, and new risks appear in those populations; for example, the population of teens with cell

and smart phones now face risks of cyberbullying (20% were victims in 2010) [10]. In light of this, identity threat modeling and risk analysis seek to reduce and manage the risk these new threats pose to different forms of identity.

Some basis for identity threat and risk modeling is found in the fields of privacy and security. Identity, privacy, and security are clearly related fields; they overlap in some areas of study, but differ in their focus. Privacy risk models often focus on the perspective of the individual rather than the device or organization. Also, in privacy risk modeling, the risk is not typically to systems or resources, but to the individual's social obligations or potential for embarrassment. Additionally, the *adversary* in privacy risk scenarios is often someone who already knows the individual's identity, e.g., a parent, employer, friend, or acquaintance [11]. The study of identity risk casts a wider net to include the study of devices, organizations, and abstract entities such as online personas, avatars, or profiles used in different systems. The identity risk analyst is interested not only in privacy risks, but also in risks to systems, resources, and reputation.

Another related field is security threat modeling [12]. Security threat modeling is an important component of identity risk modeling, but not sufficient in and of itself to model identity risk. Historically, the study of security has been about the "mechanisms and techniques that control who may use or modify the computer or the information stored in it" [13]. Modeling identity risk may mean defining or enforcing such security threat controls. However, identity threat modeling contrasts with security risk modeling in that an identity often comprises information across multiple computer systems and networks. Indeed, attributes of a single identity are often dispersed and replicated around the world, in systems both human and machine that are controlled by organizations with different interests. Traditional techniques of modeling security threats, while still useful, must be updated to handle a more distributed, more inter-connected, and more dynamic

definition of a protected resource. Security threat modeling is typically done from the perspective of a system designer who is seeking ways to prevent an attacker or intruder from accessing protected data [12]. The identity risk modeler will also consider this, but may additionally be concerned with how to prioritize incident response *after* an incident has occurred. To illustrate this, suppose a set of account names and passwords for an e-commerce web site have been breached. An identity risk modeler can prioritize which identity attributes are now at risk for the user population because she understands how the breached attributes are related to other identity attributes. For example, the breached accounts may contain users' social security numbers, and the risk model may closely associate social security numbers to credit card applications. The identity risk modeler can thus warn users to monitor their credit reports for the next six months.

IDENTITY RISK STUDIES AND REPORTS

A popular annual report is the Data Breach Investigations Report (DBIR) [14], put together each year by Verizon using data gathered from several national and international agencies, including United States Secret Service, Dutch National High Tech Crime Unit, Australian Federal Police, and others. This report is notable because it is one of the few private studies with access to large numbers of structured incidents. These incidents are shared to Verizon in a format of their design called VERIS [15], which is oriented toward understanding and managing risk. While *NewsFerret* is interested in answering many similar questions, its modeling technique differs from that used by Verizon in a couple ways. First, *NewsFerret* uses unstructured incident data (that found in news stories) while Verizon uses structured data in the VERIS format. Second, *NewsFerret* has a flexible data model which can be adjusted to model different threat

spaces, whereas Verizon is limited to modeling data breaches reported by government agencies.

The Federal Trade Commission (FTC) operates the Consumer Sentinel Network (CSN) which is a centralized database of consumer complaints. Their Bureau of Consumer Protection's Division of Privacy and Identity Protection also

> analyzes identity theft trends, promotes the development and efficacy of identity fraud prevention strategies in the financial services industry, and identifies targets for referral to criminal law enforcement [16].

CSN collects data from a number of state law enforcement agencies, and shares the information only with law enforcement groups. The FTC issues reports periodically based on this data, including one [17] that indicates "identity theft" was the top complaint category for 2011, as the cause of 279,156 complaints. *NewsFerret* contrasts with CSN reports in that *NewsFerret* uses unstructured incident data as input. Finding ways to apply *NewsFerret*'s analysis technique to large quantities of structured data such as that found in a VERIS or CSN database could be an interesting area for future study. The Future Work chapter makes note of this.

The Center for Identity's ongoing project to develop Identity Ecosystem Maps (IEMs) [2] heavily informed this work, especially in its choice to model identity ecosystems using attributes and relations. IEMs offer a framework for analysis, but do not prescribe a mechanism by which to populate values for specific attributes and relations. *NewsFerret* complements IEMs by acting as one such data input mechanism. Specifically, using *NewsFerret*, a risk analyst can draw conclusions that she then inputs to an IEM as a collection of identity attributes and weighted relations, where the weights represent semantic relatedness between two attributes.

INITIATIVES

Several recent government initiatives have motivated work in the area of identity. These include The "Homeland Security Presidential Directive 12: Policy for a common Identification Standard for Federal Employees and Contractors" [18] which established a requirement for a federal ID card in the U.S. The "National Strategy for Trusted Identities in Cyberspace (NSTIC)" established a need for updated definitions and services around identity in cyberspace [19], and a similar State Identity Credential and Access Management (SICAM) Guidance and Roadmap advocated similar approaches at the state level [20].

TEXT MINING

Latent semantic analysis (LSA) was first described in [21], and the author of this report conducted an earlier project to determine similarity of U.S. congress persons and senators based on their floor speeches in Congress [22]. The technique is a form of dimensionality reduction which exhibits an interesting property whereby an approximation of the observed data is more informational than the raw observed data, said to represent the "semantic space" representing the underlying concepts of a set of documents[23]. LSA and related techniques can be used for several purposes, including similarity comparisons in the "semantic space" and information retrieval. *NewsFerret* uses techniques for both concept similarity comparison and information retrieval. Similar techniques, sometimes referred to as "concept linkage," have been used in other domains such as systems biology[24] [4].

Chapter 3: Requirements

The Introduction chapter gave an overview of the identity risk modeling domain along with an example scenario. This section further details the functional and non-functional requirements that *NewsFerret* should satisfy in order to assist with risk modeling activities.

OBJECTIVE

The high-level objective of the system should be to provide a framework within which an *identity risk analyst* can identify and analyze risk to an identity ecosystem within a threat context through the use of a valid risk model.

STAKEHOLDERS

Users in all of the scenarios given in the Introduction may be abstracted into a user profile called the *identity risk analyst.* Responsibilities of the risk analyst include:

- **Identify risk**. The analyst identifies which elements of an identity are at risk and should be part of the risk model. For example, should "Facebook account" be part of the risk model, or is that attribute not relevant to any known threat?
- **Analyze risk**. The analyst seeks to understand the mechanisms by which the identity elements are put at risk, and seeks to quantify the likelihood that an incident should occur. To support findings, the analyst validates the risk model against existing incident data.
- **Assess risk**. The analyst values stakeholder impact—e.g., cost or damages—of the risk, should an incident occur. For example, if a user's

bank account is compromised, the cost is very high to both the bank and the customer.

- **Manage risk**. The analyst—typically in collaboration with a larger team— implements controls and countermeasures to mitigate, prevent, or counteract the risk. The analyst communicates the risk model to stakeholders, and keeps the risk model up-to-date.

To perform the above responsibilities effectively, the risk analyst must have a valid risk model. To create a valid risk model, the analyst may seek to answer the following questions:

- Given a set of compromised identity attributes within an ecosystem, what other attributes may be at risk? How much are they at risk?
- Given a set of compromised identity attributes within an ecosystem, what are the likely attack vectors, i.e., actors, actions, resources that compromised the attribute? How likely?
- Given a threat context, are the elements of my identity model valid? Are any elements missing in my model? For example, do threats exists to attributes I have not considered in my identity model?
- Given an ecosystem, are the elements of my threat model valid? Are any elements missing in my model? For example, do threats exists that I haven't considered in my threat model?

FUNCTIONAL REQUIREMENTS

The following step-by-step process illustrates the primary functional scenario the system should support to satisfy the above objectives and stakeholder

responsibilities. Notes are added to illustrate how each of these steps would be realized in the example scenario described in the Introduction chapter:

1. The risk analyst declares a set of keywords matching the threat context within which the analyst is interested in modeling risk.

   ***Example scenario**: Based on the threats under investigation, the analyst declares keywords "identity theft" and "credit card theft"*

2. The analyst defines a starter list of identity attributes that she is interested in understanding and analyzing in the above threat context. In considering this list of attributes, the analyst may consider the following attribute types, (which take after discussions on multi-factor authentication [25]):

Attribute Type	Examples
Things you *are*	Eye color, DNA, maiden name, IP address, federal tax ID, location
Things you *have*	Money, driver's license, credit card, bank account, Medicare card, brand, reputation
Things you *know*	Password, last year's tax return amount

Table 1: Identity attribute types [2]

   ***Example scenario**: Based on the starting risk model (see Figure 2), the analyst defines the starter list to include "name," "address," "e-mail," "Amazon account," and "credit card."*

3. The system gathers and stores information describing the given threat context. Typically, this information may be in the form of a description or report of a particular incident or occurrence of a threat scenario.

   ***Example scenario**: The system will gather and store news stories matching the keywords "identity theft" or "credit card theft"*

4. The system builds analytical objects that model identity risk within the threat context.

   ***Example scenario**: Analytical objects and reports are described in the Analysis chapter below.*

5. The analyst uses the analytical objects and services provided by the system to validate her existing identity risk model.

   ***Example scenario**: The analyst adds elements to the risk model based on new information in the analytical objects. A revision of the risk model described in the example scenario is described in the Analysis chapter.*

NON-FUNCTIONAL REQUIREMENTS

The system must also satisfy non-functional requirements in the following areas: privacy, scalability, and timeliness.

An important factor in the study of identity and other security-related risks is the inherent sensitivity of incident data [11]. Incidents reported to the FTC and other law enforcement agencies are not available except to other law enforcement agencies [17], since incident details often reveal personally identifiable information (PII) that could be used by a future attacker. So, the system must find some way to gather and store threat scenarios in a way that does not reveal PII. Additionally, since access to this type of data is limited and difficult to gather, the system should ideally make use of a public data source.

Second, the system should be scalable. As discussed in the introduction, manual entry of incident data into a structured format is both time consuming and error-prone. In order to get the volume of data needed to accurately portray a threat context, the system should minimize manual data entry. To support flexibility in expanding the system in the

future to gathering risk data for other domains or for specialized identity domains, the system should be able to store and process large amounts of textual data.

Finally, the system should acquire up-to-date threat data. Like other security-related risks, identity security risks often show trends over time, and the predominant threat scenarios of today may not be the predominant threats of tomorrow. For a risk model to be accurate, it must be based on timely data. As one example of a changing trend, virus attachments in spam e-mail was less of a problem in 2013 than it was 10 years earlier; however rogue links within spam e-mails increased as a threat [8]. It is important, therefore, that the risk analyst model risk based on incident data that is as up-to-date as possible. Requiring manual data entry of incident data is an inhibitor to this requirement, and so again the system should avoid requiring manual data entry where possible.

Chapter 4: Design

What follows is a description of the design of *NewsFerret*, intended as a solution to the functional and non-functional requirements described in the earlier section.

KEY HYPOTHESIS

First, an illustration of a key hypothesis of the design will explain some important design decisions of *NewsFerret*. In short, the hypothesis is that *keyword topics, e.g., "identity theft," found in news stories represent some subset of the semantic structure of identity risk*. By modeling this semantic structure and comparing it against the analyst's risk model, the analyst can reveal insights and find where her model may be valid or invalid, or may be missing elements.

News stories found in this way are not a complete or even thorough representation of the complete risk space. However, in this way, they are similar to incident reports: both are reported inconsistently, their level of detail varies, and they may be skewed toward focusing on a particular kind of threat. News stories skew their focus toward "newsworthy" events. It is difficult to quantify how this skew affects risk analysis. It may be worth mentioning that the news media generally considers identity threats to be newsworthy, especially since many threats are criminal offenses. However, quantifying the skew of news coverage is still difficult. Possibilities for future study in this area are noted in the Future Work chapter.

Despite their problems, news stories have two important qualities benefitting the design:

1. **Sheer quantity.** Media outlets publish large numbers of news stories, covering events on thousands of topics from around the world. Over a month-long period, *NewsFerret* found over 200 unique news stories on the

keyword topic of "identity theft," even when limited to an English-language U.S. locale.

2. **Public nature.** Information published in a news story is a form of public record. In using news stories, the analyst need not be overly concerned with protecting PII or incident details since the information has already been made public.

The first major design decision, therefore, is to utilize news stories as stand-ins for the real-world, semantic space of the identity risk. A description of the design that does this follows.

OPERATIONAL REFERENCE MODEL

The author begins illustrating the design using an Operational Reference Model diagram [26]. This diagram shows the high-level steps of the process that the design will realize.

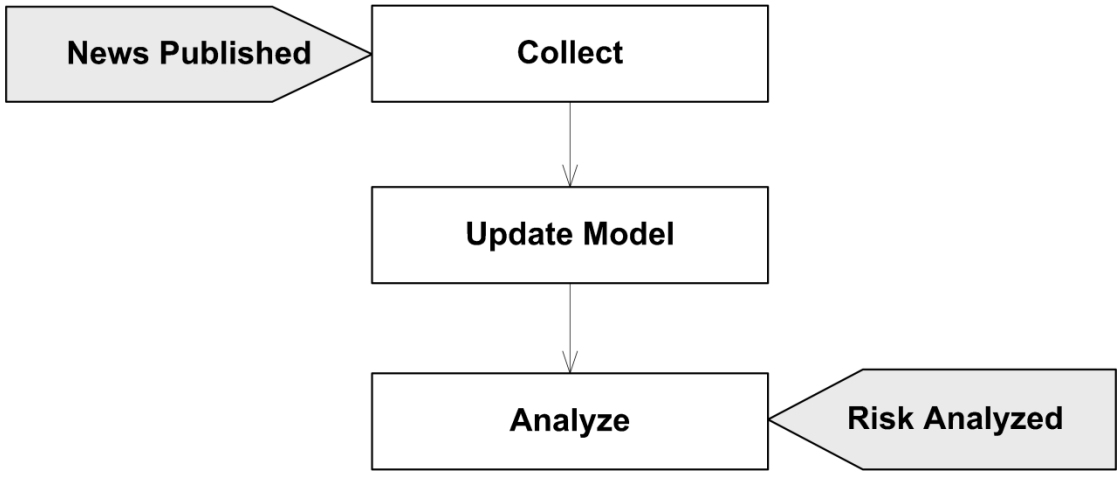

Figure 3: *NewsFerret* Operational Reference Model diagram

Elements of the diagram include the following (shown in process order):

- **News Published**. A media outlet publishes one or more news stories on a topic relevant to the domain the analyst wishes to study. For example, a regional newspaper publishes an article, "Rome woman charged with identity theft," [27] which is subsequently aggregated by a news feed publisher under the keyword topic of "identity theft."
- **Collect**. The system retrieves the news article(s), based on information from one or more news feeds; extracts the core content of the web page; and stores the news article *content* along with metadata about it, including *author*, *publish date*, *retrieval date*, *title*, and *URL*.
- **Update Model.** The system updates its analytical model (for examples, see Figures 5-8) for the domain based on all stored news stories, building analytical data structures for study and analysis.
- **Analyze.** The analyst studies the analytical data structures and validates the risk model. She exports reports and analytical objects for further study using tools such as Microsoft Excel, graph visualization tools[28], or IEMs [2]. She updates her risk model appropriately, backing unusual or unexpected conclusions with relevant news stories.
- **Risk Analyzed.** After analysis, new risks have been identified and analyzed, and the analyst has analytical objects, reports, and related news stories to support her new risk model.

COMPONENT MODEL

The design process during this project included exercises to decompose tasks from the above ORM, analyze functional dependencies, model data elements, and structure components based on the non-functional requirements. For brevity, details of

all those exercises are not presented in this report. Instead, a component model summarizes these design outputs. The diagram and descriptions below group together required functions and data elements in a way that satisfies the functional dependencies and non-functional requirements, illustrating the structure of the *NewsFerret* system from a business standpoint:

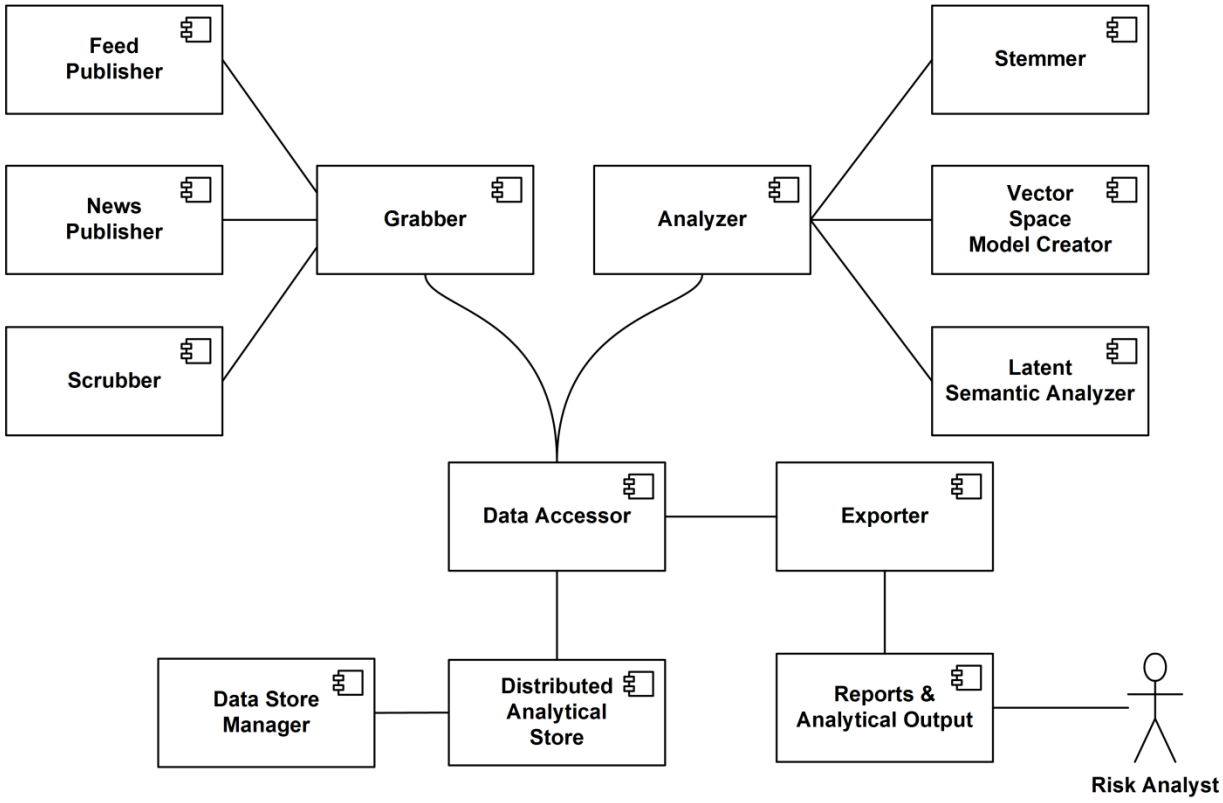

Figure 4: *NewsFerret* component model

Component descriptions are listed in operational order. Where applicable, notes are added to illustrate the component in the context of the example scenario:

- **Grabber.** Queries news feeds for news stories related to one or more keyword topics, retrieves the news stories from the News Publisher, and calls the Data Accessor to store metadata and story content.

***Example scenario**: The component retrieves and stores news stories related to "identity theft" and "credit card theft" topics.*

- **Feed Publisher.** Responds to keyword topic searches, providing a list of recent news story summaries (and URLs), with a typical date coverage range lasting anywhere from within the last 24 hours to within the last week. The Feed Publisher is a publicly available news feed reused for this project.

  ***Example scenario**: News feeds return article metadata for news related to "identity theft" and "credit card theft" topics.*

- **News Publisher.** Publishes news stories on the web for retrieval by the News Grabber. This component is provided by news media publishers.

  ***Example scenario**: News publishers provide "identity theft" and "credit card theft" news stories in HTML format.*

- **Scrubber.** Extracts the core content of a news story from its published format (typically HTML), removing ad content, menu and navigation content and so on; but retaining the title, author, and content of the article. We reuse the *boilerpipe* library for this component. [29]

  ***Example scenario**: Scrubber removes ad content and HTML tags from an HTML news article on identity theft.*

- **Data Accessor.** Provides storage-technology-independent data access functions, decoupling the news grabbing and analysis functions from the specific choice of technology for the Distributed Analytical Store.

  ***Example scenario**: Stores article content and metadata for a news story*

- **Distributed Analytical Store.** Provides storage-technology-specific data access functions for a given distributed data storage technology. Notably,

it must store the news stories and the vector space model. This component was realized by reusing the *Apache Cassandra* database [30].

- **Data Store Manager.** Storage-specific component that manages the data definitions or schema of the Distributed Analytical Store.
- **Analyzer.** Coordinates and executes the steps to create the primary analytical objects. Supports scalable distributed processing. The main steps are to start the Vector Space Model Creator to create the vector space model, and then to call the Latent Semantic Analyzer functions to create the semantic space model.

  Example scenario*: The analyzer performs the above actions against a large set of news stories on "identity theft" and "credit card theft". The resulting semantic space models contain information about related concepts within the news stories. For example, the semantic models show that the concept of "password" is very closely related to "amazon" within this set of news stories.*
- **Stemmer.** Reduces words to a stem form in order to reduce the number of terms with similar meanings. For example, "bank," "banks," and "banking" will all reduce to the same stem. Part of this component reuses a publicly available Java implementation of the Porter stemming algorithm [31].
- **Vector Space Model Creator.** Creates a term-document matrix from the complete set of news stories. A term-document matrix is an matrix where each element *(i,j)* represents the number of times term *i* occurs in document *j*.

- **Latent Semantic Analyzer.** Implements the key algorithm to create a reduced-rank singular value decomposition (SVD) representing a model of the "semantic space" of all news stories in a threat context. Should operate in a distributed fashion.[1] This component reuses several MATLAB [32] built-in functions to realize the algorithm.
- **Exporter.** Exports reports and analytical objects for use by the risk analyst in validating her risk model. The Analysis section describes sample reports and objects.
- **Reports & Analytical Output.** Reports and data files used directly by the risk analyst and imported into other tools in order to validate a risk model. Example reports and output may be found in the Analysis chapter.

TECHNOLOGY SOLUTIONS

The system implements the above components in the Java and MATLAB programming languages. Choice of language was driven by development familiarity with these two languages, and by availability of open source libraries in these languages to assist with various functions, e.g., scrubbing [29], feed parsing [33], and linear algebra functions [32].

Due to the potential amount of news story content and the scalability requirements, the core processing and data storage components require a scalable technology, and they must also work well together. While not evaluated in-depth for performance characteristics in this report, the system used Apache Hadoop for distributed map-reduce style processing [34], and Apache Cassandra for write-fast, horizontally scalable data storage [30]. Use of this style of storage and processing lends flexibility to

[1] This project does not implement distributed LSA processing

expand *NewsFerret* in the future to gather large quantities of news stories across a variety of specialized threat contexts, and still continue to provide near-time analysis.

For ease of interoperability with other analytical tools including MATLAB, Gephi (for graph visualization), and spreadsheet applications, the system exports analytical objects in comma-separated value (CSV) or whitespace-delimited format (DAT).

SOURCE CODE

The author of this report published all source code, along with installation, build, and configuration instructions to a publicly accessible GitHub repository [35].

Chapter 5: Analysis

The author implemented the above design and ran the system—as might an identity risk analyst—over a period of several weeks. What follows in this chapter are results of the run of the completed *NewsFerret* system along with a description of how the risk model from the example scenario (see Introduction chapter) could be validated and revised based on such results.

INPUTS

The first step in using *NewsFerret* is to define a threat context by defining a news feed and keyword topic to use in searching for news stories. No exact process for this is prescribed by the author, although a trial-and-error process with some validation is suggested. After some selective sampling from different news feeds, two news feeds with keyword-based URLs for "identity theft" were selected to represent the threat context:

Feed Publisher	**URL**
Google News	http://news.google.com/news?um=1&ned=us&hl=en&q=identity+theft&output=rss
Huffington Post	http://www.huffingtonpost.com/tag/identity-theft/feed

Table 2: News feeds and keyword-based URLs for an "identity theft" threat context

Additionally, the analyst inputs a list of identity attributes to represent the portion of the identity ecosystem to analyze. The Identity Ecosystem Maps project at the Center for Identity has produced a list of identity attributes [36]. To this list were added the

attributes from the starting risk model (see Figure 1). A text document containing the resulting list of attribute terms was input to *NewsFerret*.

DURATION

NewsFerret's Grabber component ran at least once per day for five weeks between mid-February and late-March 2013. During that period, the system amassed 215 news feed listings, and 206 valid news stories (some URLs failed to return valid stories).

Although not implemented, a daily scheduled job would seem appropriate if one were to run the Grabber for an extended period for multiple news topics.

ANALYTICAL OUTPUT

Next, the analyst performs steps necessary to create a set of analytical objects. At any point after news grabbing has begun, the analyst runs the Analyzer component. The analyst can then export the resulting analytical objects. Although results can be retrieved at any time, results improved over time as additional stories were gathered. Patterns started to emerge and stabilize after around 100 news stories were accrued (taking around 2.5 weeks). The primary analytical output comprises two CSV files:

1. Pairwise semantic relatedness measures for important concepts
2. Top 5 most related news stories for each pair of important concepts

Concept Relatedness

Below is an example snippet of the pairwise relatedness measures, as viewed in a Microsoft Excel spreadsheet:

	A	B	C	D
1	source	target	weight	type
251	amazon	accounts	0.3285	undirected
252	amazon	banking	0.4195	undirected
253	amazon	user	0.4366	undirected
254	amazon	email	0.4411	undirected
255	amazon	unique	0.4917	undirected
256	amazon	shopping	0.507	undirected
257	amazon	website	0.5671	undirected
258	amazon	twitter	0.6416	undirected
259	amazon	group	0.6474	undirected
260	amazon	passwords	0.6686	undirected
261	american	female	0.2099	undirected
262	american	proof	0.2937	undirected
263	american	mortgage	0.3691	undirected
264	american	color	0.4045	undirected
265	american	ebay	0.4463	undirected
266	american	travel	0.5868	undirected
267	apple	ship	0.2009	undirected
268	apple	transunion	0.2146	undirected
269	apple	equifax	0.2146	undirected
270	apple	store	0.3108	undirected
271	apple	express	0.4302	undirected
272	apple	design	0.8141	undirected
273	apple	computers	0.8494	undirected
274	apple	ipad	0.9548	undirected

Figure 5: Example report of relatedness measures for important term pairs.

Legend: “Source” and “Target” columns contain the two important concepts; “Weight” column contains the measure of relatedness between the two concepts, using a continuous scale between 0 and 1, where 0 means the two concepts are not related at all, and 1 means the concepts are very closely related. The “Type” column is for use by graph visualization tools, but also demonstrates that the relation between the concepts is not directed. Highlighted cells indicate relations with a high degree of semantic relatedness

Next, the relatedness measures were fed into a graph visualization tool for further analysis. The relatedness measures are useful to examine within a spreadsheet, but a

graph visualization gives a powerful high-level picture of the data. Below is a visualization of the same information using the graphing tool Gephi:

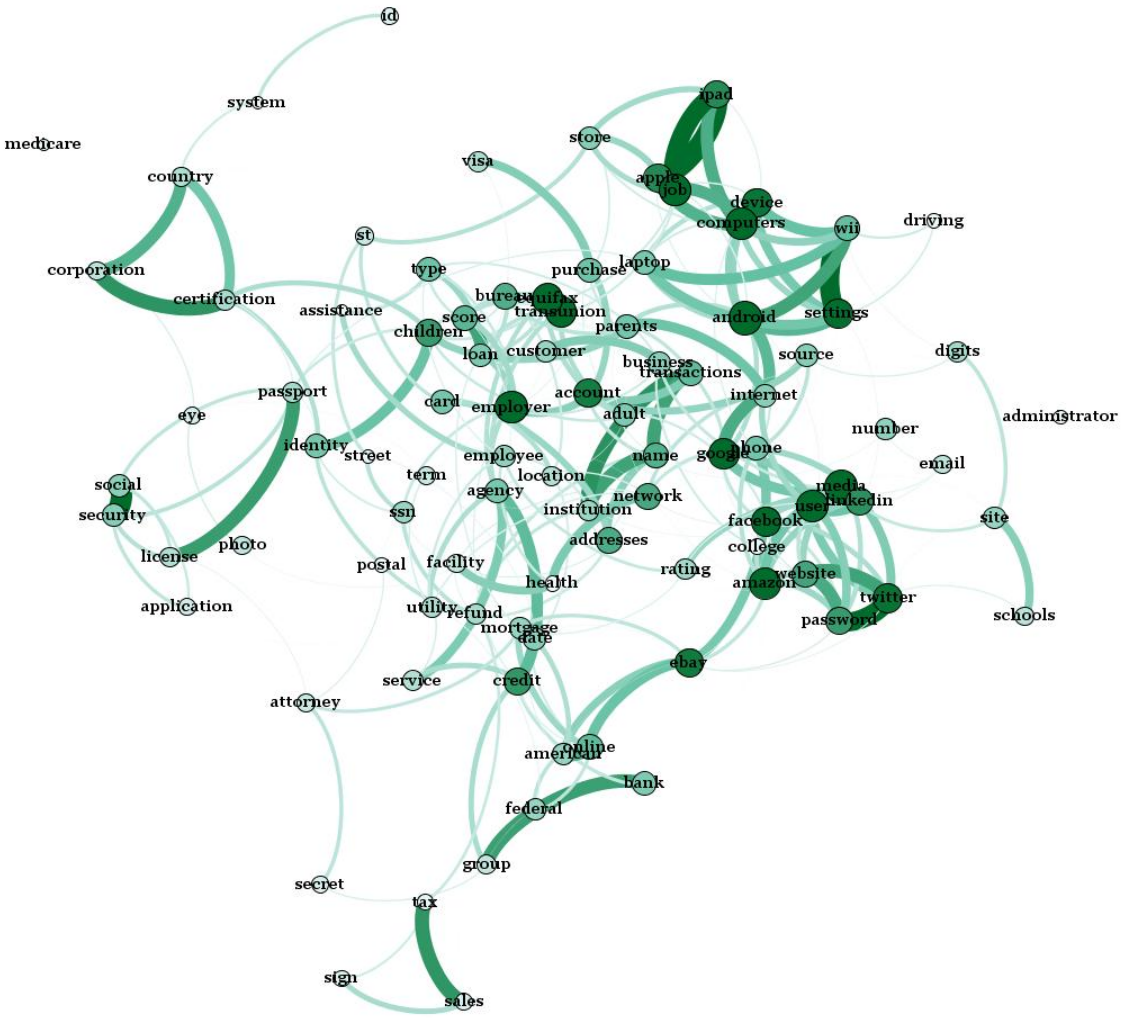

Figure 6: Graph visualization of concept relatedness and importance

Legend: Concepts are represented by labeled nodes, and relatedness between concepts represented by edges. The thicker the edge, the more related the two concepts. The larger the node, the higher the weighted in-degree of the node—a rough measure of the importance of that concept to the identity risk context. "amazon," "facebook," "credit," and "android" display high levels of importance to the identity risk context, while curiously strong relatedness is exhibited between "twitter" and "password."

In exploring Figure 6, the risk analyst may identify specific relations that are unexpectedly strong. The reason for such relationships is not always clear. To further explore, the graph visualization tool can help by only showing nodes in direct contact with a specific selected node of interest. For example:

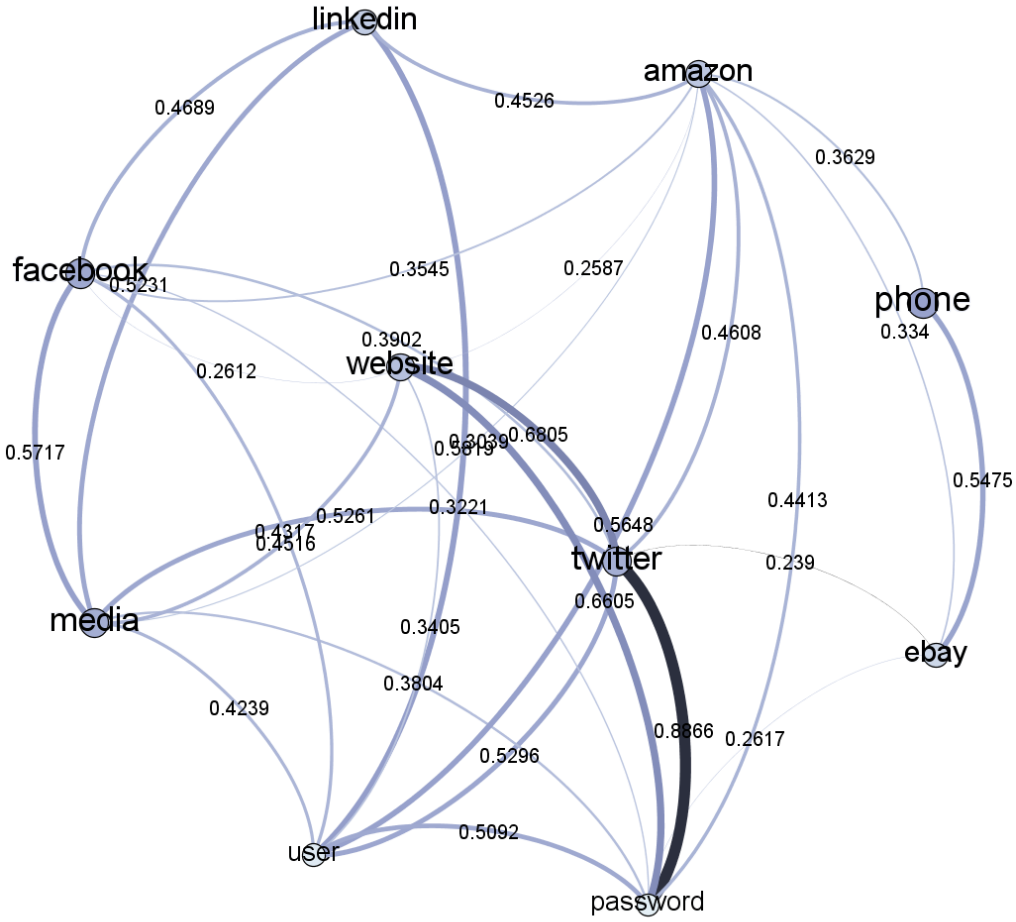

Figure 7: Concepts closely related to "amazon"

This graph shows all concepts closely related to "amazon" by a measure of 0.2 or higher. Some unexplained relationships deserving further exploration might be those between "amazon" and "twitter," "amazon" and "password," "amazon" and "phone."

And another example that looks at a topical concept:

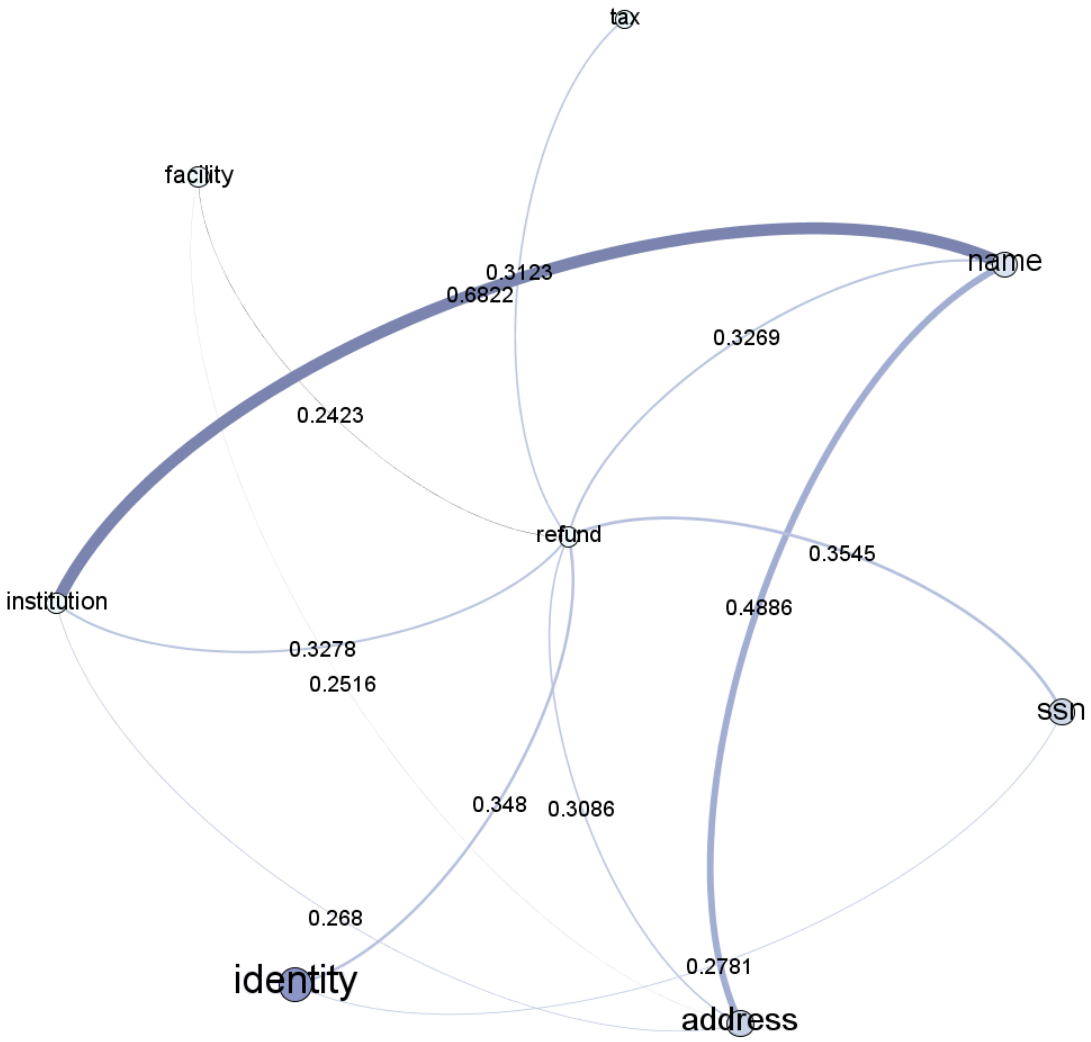

Figure 8: Concepts closely related to "refund"

The news gathering was run during tax season, so several stories focused on a recent trend of IRS tax fraud. Figure 7 shows the concepts related to refund by a measure of 0.2 or higher. One might infer from this that "address," "name," "ssn," are all attributes that are involved in this particular threat scenario, although further research and interpretation would need to be done to confirm.

The risk analyst now must now interpret the analytical models to validate the starting risk model. The relation between "amazon" and "password" concepts, for example, may mean that Amazon passwords are being hacked, or may mean that that a large breach of passwords occurred recently; or it could mean a host of other things. The analyst has an additional tool to help interpret the above graphs in a way that can validate her model.

Related News Stories

The relatedness measures shown above for pairs of concepts indicate the level of meaningfulness in the relationship between two concepts in the context of "identity theft" (the threat context). But what does that mean for any two specific concepts? *NewsFerret*'s analytical models do not state this explicitly and do not represent cause-and-effect relationships. What they do indicate is *whether the two concepts often appear in the same news stories related to the given threat topic*. To further explore why this might be, one suggests two techniques for the risk analyst to analyze and explain the relatedness.

First, one offers some heuristic guidance. Experience in examining multiple pairs of results showed a pattern that if the relatedness measure was above .9, then the relationship was likely one of synonymy, or else the concepts were parts of a multi-word term[2]. If the relatedness measure was below 0.2, then generally there was little to no meaningful relationship between the two concepts in this threat context. Typically, the relationships of interest to the risk modeler lay between these two extremes.

The heuristic technique is only so helpful. A large range of related concepts remain for the analyst to explore. The analyst still wants to understand why each of the

[2] Although not implemented in this work, the Future Work chapter suggests n-gram analysis to more gracefully handle multi-word concepts, e.g., "Social Security Number," "Credit Card," etc.

remaining term pairs are related at the level they are. For example, the terms "American" and "travel" receive a high relatedness measure (0.5868) as shown in Figure 5. Why is this? To explore this, the system repurposes the semantic model as a search engine, and conducts a search for the top five most related documents for each concept pair. Here is a sample of such search results:

Concept	Concept	Relatedness	Top 2 Related Story Titles
amazon	password	0.6686	"Identity Theft -- Your Use of Passwords Could Be Your Only Line of Defense" "The Most Popular Online Password Is 'Password,' Yet We Blame Our Privacy Problems on Facebook"
american	travel	0.5868	"U.S. Busts Massive Fake ID Scheme" "Here's Where You're Most Likely To Get Scammed"
birth	passport	0.8617	"Former Lower South man facing identity theft charges here, now in trouble with ..." "5-year-olds victims of identity theft"
controls	profile	0.5721	"Facebook users risk identity theft, says famous ex-conman" "Networkers: LinkedIn and Google+ Users Have a Higher Incidence of Identity Fraud"

Table 3: Analyzing concept relatedness through related news stories

By exploring the titles and content of the news stories related to the concept pairing, the analyst can seek support to explain why the semantic model finds two concepts to be closely related. For example, one might infer from the table above that "amazon" and "password" are related in the risk model because there have been incidents or alerts about Amazon accounts being compromised due to weak passwords. One may also surmise that "birth" and "passport" are related because falsified birth certificates have been used by identity thieves to acquire passports.

REVISITING THE EXAMPLE SCENARIO

Equipped with these new analytical models, this section revisits the example scenario and risk model described in the Introduction chapter, and describes how the analyst from that scenario would validate and revise the starting risk model.

Based on the concept relatedness graph from Figure 7, the analyst has identified that—somewhat surprisingly—the name, address, and e-mail identity attributes are not as closely related to "amazon" within this threat context as previously theorized. Instead, the "password" identity attribute appears to be at greater risk due to the higher edge weight between "amazon" and "password."

The edge weight alone is not sufficient to make a conclusion, however. So the analyst next researches the related articles from Table 3. Based on a reading of these news stories, a conclusion is reached that "password" is indeed at higher risk. Similar conclusions are reached for "amazon's" relations to the "twitter" and "linkedin" concepts. Based on these conclusions, the analyst isolates the relevant portions of Figure 7 into a sub-graph. This sub-graph is shown in Figure 9. Note that "name," "address," and "e-mail" attributes are not represented in Figure 9 because the edge weight between these attributes and "amazon" was very low (< 0.1):

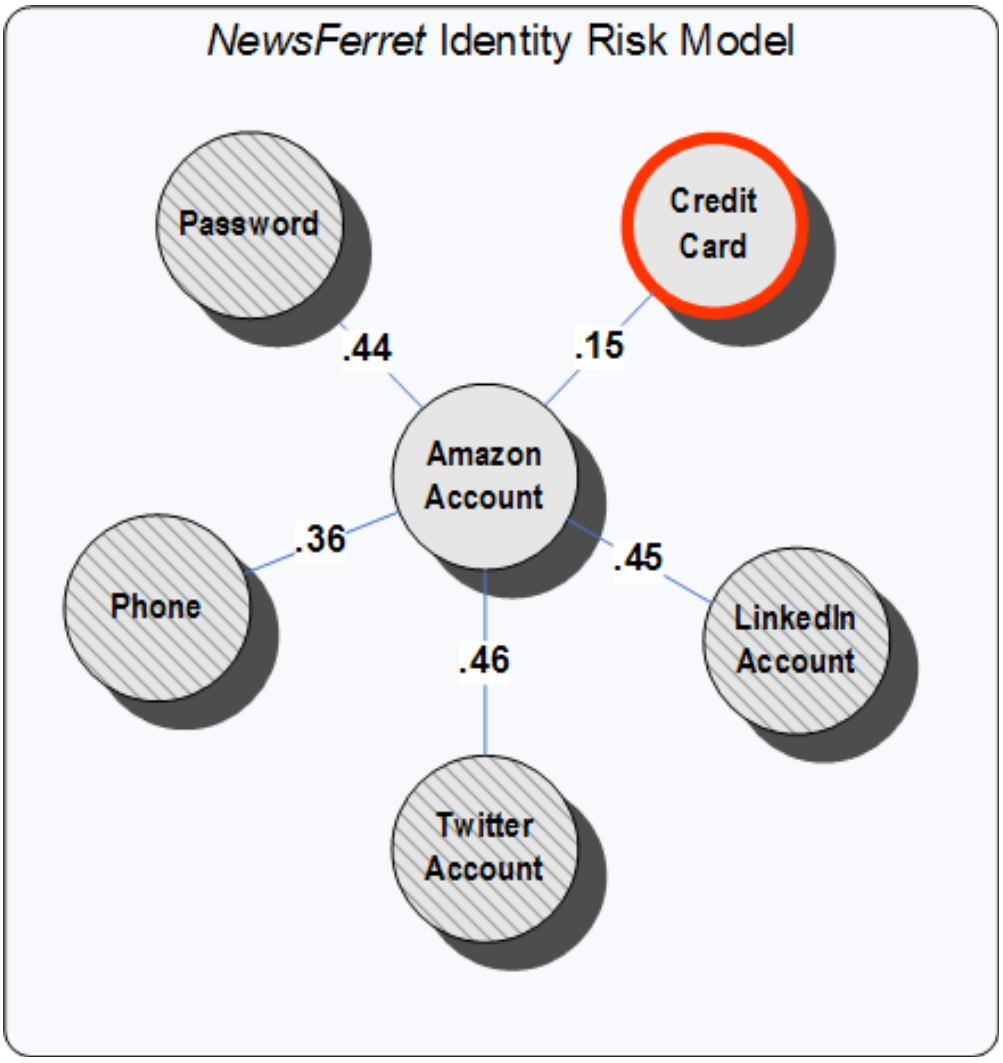

Figure 9: Sub-graph of *NewsFerret* model relevant to example scenario

Based on Figure 9, the analyst determines that updates are needed to the original risk model. After discussion and evaluation with the client, a decision is made to leave in the original attributes of "name," "address," etc., in the model, and add the newly discovered attributes. The final risk model then looks like this:

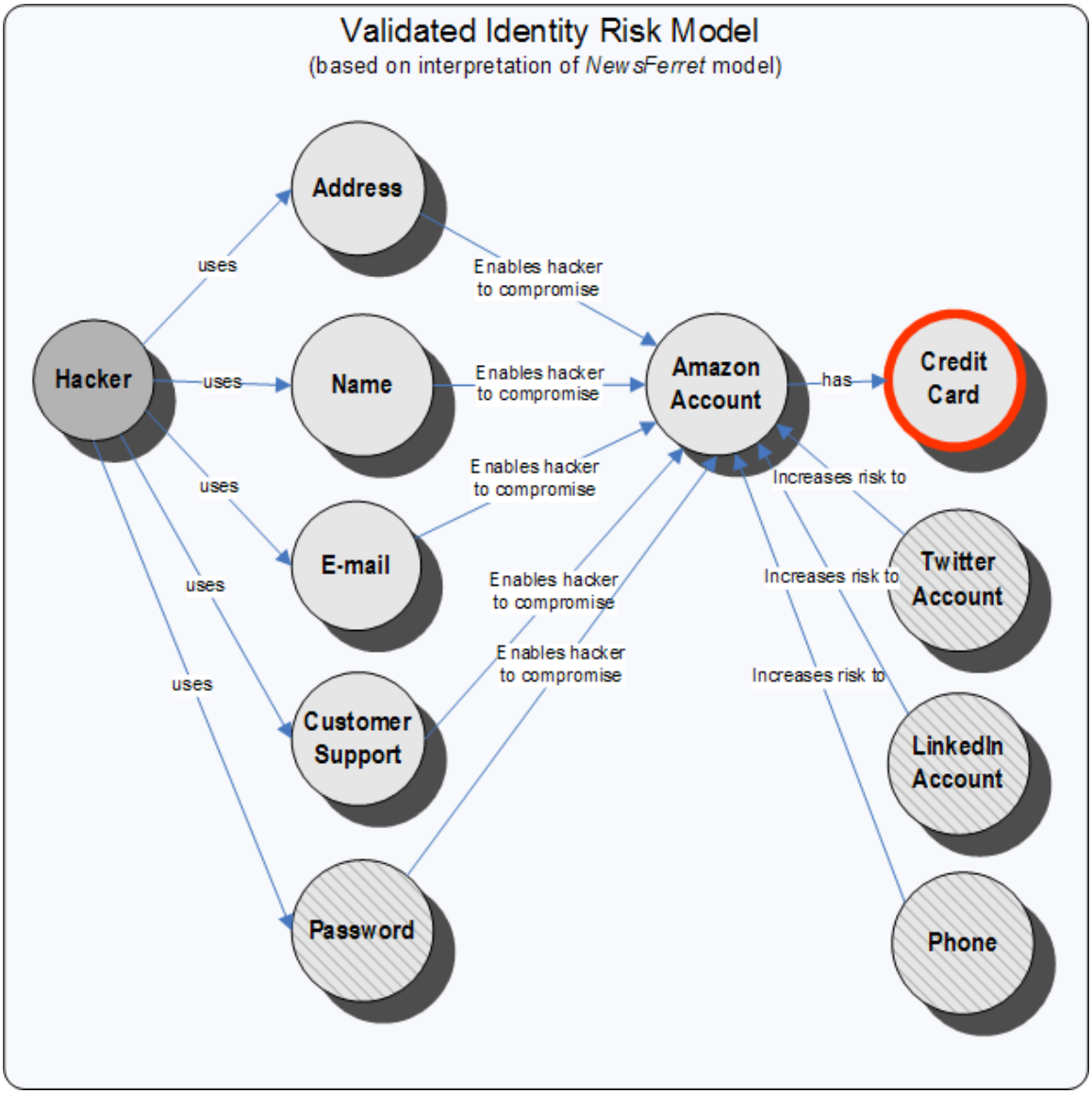

Figure 10: Validated identity risk model from example scenario.

Components with slanted lines indicate identity attributes which have been added to the original risk model based on an interpretation of the *NewsFerret* analytical model.

Using the model represented in Figure 10, the analyst is now able to answer the client's original questions (refer to the Example Scenario section) with the following information:

- The likeliest attack vectors are "password," "twitter," or "linkedin."

- Elements missing from the original model were “password,” “twitter,” “linkedin,” and “phone.”
- Passwords seem to pose the greatest risk. Effort should be expended to improve password practices or authentication techniques. Cautions should be issued to users about what information they expose on their social media accounts, and smart phone users should be advised to exercise special care to protect their phones and Amazon accounts in the event that their phones are stolen. These recommendations are formulated based on a careful reading and interpretation of relevant news stories returned by *NewsFerret*.

Chapter 6: Conclusion

This report has outlined the requirements of the identity risk analyst, explaining the need to validate identity, threat and risk models against the true semantic structure of the identity ecosystem, threat context, and identity risk. It situated the study of this problem in the context of existing threat modeling research and known text mining approaches for studying semantic structure of documents. The report then walked through the conceptual framework and functional requirements for a system that would be able to provide some validation of identity risk models, and illustrated a design that is able to satisfy those requirements by gathering and analyzing news stories as representatives of the threat context under analysis. Along the way, the report introduced and periodically revisited an example identity risk modeling scenario in order to demonstrate a sample risk model and its evolution using *NewsFerret*.

The author of this report implemented the design, and demonstrated the process to configure and run the system, resulting in analytical output from a month-long run that gathered over 200 unique news stories on the keyword topic of "identity theft." In demonstrating the analytical output and its effect on the example scenario, this report showed that concept relatedness measures can reveal unexpected relations between concepts in a risk model, and that further support for those relations can be found by exploring the news stories related to the pair of concepts.

Chapter 7: Future Work

The author of this report suggests the following future work related to this study. Since news articles take around 2.5 weeks to accrue to a reasonable level of usefulness, it would be preferred by the casual risk analyst that a set of pre-configured threat contexts be available. The system admin could organize these preset threat contexts by threat type or by some other facet such as industry vertical (e.g., health care fraud, financial fraud).

One text mining issue remained unsolved during this study, though solutions no doubt abound: multi-word concepts were handled somewhat inelegantly. For example "social security number" was treated as three separate concepts. Ultimately, this did not greatly affect the results, since the terms would end up labeling concepts that were very closely related in the graphs. However, this could be handled better—perhaps by n-gram analysis—and perhaps improve interpretability.

The author does not overlook the fact that news stories reflect a trendiness bias. Further study could be done to study the modeling bias news stories on identity threats and other threat contexts. This might improve interpretability.

The Related Work section noted some differences between security threat modeling and identity risk modeling in that the latter must often model a more distributed, inter-connected set of protected resources. While this distinction is useful, the fields of identity and security remain closely related, and NewsFerret could be adapted to model and validate cyber-security risk should it be configured with the right set of keyword topics and a set of security (instead of identity) attributes.

Finally, it should be considered an open area of research whether a system similar to *NewsFerret* could work with more structured incident reports, e.g., reports similar to those gathered by FTC or Verizon. Such a study could confirm whether LSA techniques

could enable forms of analysis on this data without requiring the costly and error-prone manual data entry.

References

[1] Javelin Strategy & Research, "2012 Identity Fraud Report: Social Media and Mobile Forming the New Fraud Frontier," Javelin Strategy & Research, Pleasanton, 2012.

[2] Center for Identity, "Mapping of the Identity Ecosystem," Technical Report TR-010413-2013, 2013.

[3] Mat Honan, "How Apple and Amazon Security Flaws Led to My Epic Hacking," *Wired*, August 2012.

[4] Weigo Fan, Linda Wallace, Stephanie Rich, and Zhonju Zhang, "Tapping the Power of Text Mining," *Communications of the ACM*, vol. 49, no. 9, pp. 77-82, September 2006.

[5] Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman, "Indexing by Latent Semantic Analysis.," *Journal of the American Society for Information Science*, pp. 391-407, 1990.

[6] Aaron Smith, "46% of American adults are smartphone owners," Pew Research Center, Washington, D.C., 2012. [Online]. http://pewinternet.org/Reports/2012/Smartphone-Update-2012.aspx

[7] socialbakers. (2013, March) socialbakers. [Online]. http://www.socialbakers.com/facebook-statistics/united-states

[8] Cisco, "2013 Cisco Annual Security Report," 2013. [Online]. https://www.cisco.com/web/offer/gist_ty2_asset/Cisco_2013_ASR.pdf

[9] Kathryn Zickhur, "Three-quarters of smartphone owners use location-based services," Pew Internet, Washington, D.C., 2012. [Online]. http://pewinternet.org/Reports/2012/Location-based-services.aspx

[10] Sameer Hinduja and Justin W. Patchin, "Bullying, Cyberbullying, and Sexual Orientation," Cyberbullying Research Center, Fact Sheet 2011. [Online]. http://www.cyberbullying.us/cyberbullying_sexual_orientation_fact_sheet.pdf

[11] Jason I. Hong, Jennifer D. Ng, Scott Lederer, and James A. Landay, "Privacy risk models for designing privacy-sensitive ubiquitous computing systems," in *Proceedings of the 5th conference on Designing interactive systems: processes, practices, methods, and techniques*, Cambridge, MA, 2004, pp. 91-100.

[12] Shawn Hernan, Scott Lambert, Tomasz Ostwald, and Adam Shostack, *Uncover Security Design Flaws Using The STRIDE Approach*, 2006th ed.: MSDN Magazine, 2006. [Online]. http://msdn.microsoft.com/en-us/magazine/cc163519.aspx

[13] Jerome H. Saltzer and Michael D. Schroeder, "The protection of information in computer systems.," in *Proceedings of the IEEE*, vol. 63.9, 1975, pp. 1278-1308.

[14] Verizon, "2012 Data Breach Investigations Report," Verizon, Corporate Report 2012. [Online]. http://www.verizonbusiness.com/resources/reports/rp_data-breach-investigations-report-2012_en_xg.pdf

[15] Verizon. (2012, August) VERIS Community. [Online]. http://www.veriscommunity.net

[16] Federal Trade Commission. (2013, March) Division of Privacy and Identity Protection. [Online]. http://ftc.gov/bcp/bcppip.shtm

[17] Federal Trade Commission, "Consumer Sentinel Network Data Book for January-December 2011," Federal Trade Commission, 2012. [Online]. http://ftc.gov/sentinel/reports/sentinel-annual-reports/sentinel-cy2011.pdf

[18] Department of Homeland Security, "Homeland Security Presidential Directive 12: Policy for a common Identification Standard for Federal Employees and Contractors," 2004. [Online]. http://www.dhs.gov/homeland-security-presidential-directive-12

[19] "National Strategy for Trusted Identities in Cyberspace: Enhancing Online Choice, Efficiency, Security, and Privacy," White House, 2011. [Online]. http://www.whitehouse.gov/sites/default/files/rss_viewer/NSTICstrategy_041511.pdf

[20] NASCIO, "State Identity Credential and Access Management (SICAM) Guidance and Roadmap," NASCIO, 2012.

[21] S. Deerwester, S.T. Dumais, and G., Landauer, T.K., Harshman, R. Furnas, "Indexing by latent semantic analysis," *Journal of the American Society for Information Science*, vol. 41, pp. 391-407, 1990.

[22] Ryan Golden, Ashton Mozano, Yousif Seedham, and Fahd Siddiqui, Mining Capitol Hill Speeches, 2010, Course assignment for EE380L: Data Mining, Spring 2010.

[23] Pang-Ning Tan, Michael Steinbach, and Vipin Kumar, *Introduction to Data Mining*.: Pearson Education, Inc., 2006.

[24] Sophia Ananiadou, Douglas Kell, and Jun-ichi Tsujii, "Text mining and its potential applications in systems biology," *Trends in Biotechnology*, vol. 24, no. 12, pp. 571-579, 2006.

[25] Lawrence O'Gorman, "Comparing Passwords, Tokens, and Biometrics for User Authentication," *Proceedings of the IEEE*, vol. 91, no. 12, pp. 2021-2040, 2003.

[26] Suzanne Barber, Course lecture notes for EE382C: Requirements Engineering, Spring 2011, 2011.

[27] Observer-Dispatch. (2013, March) UTICAOD.com Observer-Dispatch: The Mohawk Valley's Information Source. [Online]. http://www.uticaod.com/news/x1551260940/Rome-woman-charged-with-identity-theft

[28] Gephi. (2013, March) Gephi. [Online]. http://gephi.org/

[29] boilerpipe. (2013, March) boilerpipe: Boilerplate Removal and Fulltext Extraction from HTML pages. [Online]. https://code.google.com/p/boilerpipe/

[30] Eben Hewitt, *Cassandra: The Definitive Guide*. Sebastopol, CA: O'Reilly Media, 2010.

[31] M. F. Porter, "An algorithm for suffix stripping," *Program*, vol. 14, no. 3, pp. 130-137, 19080.

[32] MathWorks. (2013, March) MATLAB: The Language of Technical Computing. [Online]. http://www.mathworks.com/products/matlab/

[33] Confluence. (2013, March) ROME. [Online]. https://rometools.jira.com/wiki/display/ROME/Home

[34] Apache Hadoop. (2013, March) Apache Hadoop. [Online]. http://hadoop.apache.org/

[35] Ryan Golden. (2013, March) NewsFerret. [Online]. https://github.com/ryancgolden/NewsFerret

[36] Brian Soeder and Suzanne Barber, Center for Identity Baseline Data, 2013, Internal Memo.

[37] Phillip J. Windley, *Digital Identity*.: O'Reilly Media, Inc., 2005.

[38] Facebook, "Facebook, Inc.," Form S-1 Registration Statement 2012. [Online]. http://www.sec.gov/Archives/edgar/data/1326801/000119312512034517/d287954ds1.htm

[39] InfoGraphic Labs. (2012, February) Infographic Labs. [Online]. http://infographiclabs.com/wp-content/uploads/2012/02/fb2012_960px.jpg

[40] "2011 American Community Survey 1-Year Estimates," U.S. Census Bureau,. [Online]. http://factfinder2.census.gov/faces/tableservices/jsf/pages/productview.xhtml?pid=ACS_11_1YR_S0101&prodType=table

[41] The World Bank, "Data Bank," The World Bank, 2012. [Online]. http://data.worldbank.org/

[42] Center for Identity. (2013) Center for Identity. [Online]. http://identity.utexas.edu/research/projects/identity-threat-assessment

[43] Victoria J. Rideout, Ulla G. Foehr, and Donald F. Roberts, "Generation M2. Media in the Lifes of 8-to 18-Years-Olds. A Kaiser Family Foundation Study.," Henry J. Kaiser Family Foundation, Menlo Park, California, 2010. [Online]. http://www.kff.org/entmedia/upload/8010.pdf

[44] M. and Lipner, S. Howard, "The Security Development Lifecycle," 2006. [Online]. http://www.microsoft.com/security/sdl